DOCUMENT RESUME

ED 395 997 TM 025 196

AUTHOR McLarty, Joyce R.; And Others
TITLE Listening and Writing: Two Skills for the Price of One?
PUB DATE Apr 96
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Assessment; High Schools; *High School Students; *Listening Comprehension Tests; Pilot Projects; *Scoring; Simulation; *Test Construction; Test Use; Vocational Education; *Writing Tests
IDENTIFIERS *Work Keys (ACT)

ABSTRACT

The Work Keys (trademark) system of the American College Testing program has combined assessments of listening and writing into a single test administration. The Listening and Writing assessment uses independent scorings of the examinee's written responses to audiotaped prompts to assess both listening and writing skills separately. This paper presents the reasons test developers chose to combine the skill assessments, describes the way in which they were combined, presents some results of that choice to date, and discusses their implications for the future. The primary reasons for combining these skills were that it saves testing time for examinees and provides a better simulation of the ways the skills are used in the workplace. Test development began with drafts that were revised based on expert advice and pilot testing. Scorer training methods and administration procedures were also developed with pilot tests. The final form was studied in Wyoming with 12 test sites and a sample of 256 vocational education students and later tested with 7,097 examinees. Results of these studies supported the combination of the assessments and the use of the Work Keys Listening and Writing Assessment. An appendix contains the skill scales. (Contains four tables and eight references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 395 997

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Joyce R. McLarty
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Listening and Writing: Two Skills for the Price of One?

Joyce R. McLarty, Ph.D.
Director, Work Keys Development
American College Testing

Xiaohong Gao, Ph.D.
Psychometrician, Research
American College Testing

Marcia K. Stientjes
Test Specialist II, Work Keys Development
American College Testing

In G. Englehard (Chair), Assessment in the Humanities. A symposium to be presented at the National Council of Measurement in Education 1996 Annual Meeting, New York City.

TM025196

BEST COPY AVAILABLE

Listening and Writing: Two Skills for the Price of One?

The advent of performance assessment has sparked increased interest in the integrated assessment of multiple skills. Using the same examinee performance to assess several skills is seen as desirable for two important reasons: it can reduce testing time by requiring only a single response that indexes multiple skills, and often it can better model the real world in which one routinely uses multiple skills interactively. The second reason is especially important when a significant part of the purpose of assessment is to motivate learning. However, the combination of multiple skills into a single assessment has its challenges as well as its benefits. In constructing such an assessment, attention must be given to both the unique and the joint contributions of each skill in order to ensure that examinees responding to the assessment demonstrate all of the skills being measured. Interactions by which the performance of one skill can augment or detract from that of another are especially important to consider in the assessment development process. If separate scores are to be reported for the individual skills, scoring keys must isolate their independent contributions to the response, and careful attention must be given to the possibility that bias in scoring one skill may be introduced by the examinee's demonstration of another.

ACT's Work Keys™ system has combined the assessments of listening and writing into a single test administration. The *Listening and Writing* assessment uses independent scorings of the examinee's written responses to audiotaped prompts to separately assess both listening and writing skills. Because only two, fairly clearly identified skills are evaluated in this assessment, the Work Keys *Listening and Writing* assessment facilitates an examination of the effects of each on the other in the combined assessment, thus providing a clear view of the potential effects of skill combinations more generally. This paper presents the reasons for choosing to combine these skill assessments into a single administration, a description of the way in which these skills were combined, the results of that choice to date, and the implications of those results for the future.

I. Reasons for combining Listening and Writing into one assessment for the Work Keys system

The primary reasons for considering combining skill areas for the Work Keys system were the same ones that generally apply: it saves testing time for examinees, and it provides an opportunity to better model the way these skills are used in the workplace. Workplace writing is commonly based on some type of externally provided information, usually from written or spoken sources. While some workplace writing requires individual research and personal opinion, much of it is simply the transmission of information from a source to a designated recipient. In order to assess writing skills at a fundamental level (higher level composition skills were not included in the assessment specifications), focusing on transmission of information appeared a strong choice.

Choosing to use the transmission of information as the basis for the writing assessment as contrasted with, for example, the composition of a personal essay, reduced the effect of personal background and experience on assessment performance. Choosing to provide the information in spoken (audiotaped) rather than written form, reduced the likelihood that examinees would simply copy the spelling, sentence structure and other syntactical features from the stimulus materials provided, which could possibly invalidate some aspects of the writing score. It also reduced the impact of examinee reading skills on the results of the assessment and facilitated rater attention to the technical quality of the writing with minimal interference from the content of what the examinees had written.

Since listening is essentially an internal process and determining whether someone has correctly heard the material is not possible unless some response has been made, the assessment of listening skills required the choice of a response format. Three alternatives were considered: performance, a written response, and a selected response. Performance (doing what one was told to do) was a strong model for many workplace situations, but was limited in that many things listened to do not lend themselves to a performance response. A selected response approach (e.g., multiple choice) was feasible but might have favored recognition of the content rather than recall of the information. Use of a constructed response format requires examinees to retain the information and also produce a response that can then readily be scored for accuracy and completeness. Based on these considerations, combining listening and writing skills assessments for Work Keys seemed both feasible and desirable.

II. Development of the Work Keys *Listening and Writing* assessment

A. Prototyping

Development of the Work Keys Listening and Writing skill areas began with the study of existing writing and listening assessments, augmented by a review of the relevant instructional and research literature in both skill areas. A general description of each skill area and of the assessment approach, sample items, and a scoring guide for the proposed assessment were developed and revised several times in the fall of 1991. The assessment gradually evolved from three levels of difficulty to four. Levels of difficulty (or complexity) were determined by the number of pieces of information contained in a message, with the range extending from 3-4 essential pieces plus 2-3 non-essential pieces at the lowest level to 11-13 essential pieces plus 4-6 non-essential pieces at the highest level. These complexity/difficulty levels are distinct from the skill levels attached to examinee performance and discussed later in this paper. Throughout the various revisions, the focus was on workplace messages and it was maintained that accuracy of content was of primary importance in the evaluation of listening skills and would not be considered in judging the writing quality. The current descriptions of the Listening and Writing skill scales are provided in the Appendix.

A local pre-prototype administration of the assessment was conducted with approximately 15 examinees. This was a short version with only four prompts (one at each level) that lasted approximately 30 minutes. Comments recorded by staff after that administration indicated

that many examinees did not write responses in complete sentences, but did record a reasonable portion of the information. Writing in the third person was apparently difficult for many examinees and trouble following the directions was also noted. Revisions were made based on this early feedback.

A second prototype called *Listening and Composing Written Communications*, containing 12 messages, each repeated twice, with a total testing time of approximately 90 minutes was then developed. That prototype was administered in Kenosha, WI, and in Memphis, TN, by Work Keys development team members. At each site, both employees and secondary school students were tested, and they, as well as their employers and teachers/instructors, were provided an opportunity to review and comment on the test. Comments included such things as complaints of hand/finger fatigue and suggestions of numerous timing changes including the possibility of a break midway through the test, a greater length of time allowed for writing, and less time between repetitions of the message. One business representative mentioned that he was concerned that mixing the assessment of listening and writing might inhibit people who were good listeners but poor writers from showing their true listening ability.

The development team met to revise the directions, and modify the format (e.g., adding a bell to alert examinees to the end of writing periods and the beginning of new information, shortening pauses between repetitions, and adding editing time at the end of the test), and to adjust level specifications (e.g., not requiring complete sentences at the lowest level, and modifying the top level to include information given by two speakers). The prototype was re-designed to include these modifications, resulting in 12 messages with an approximate testing time of 75 minutes. The assessment was also re-named *Listening and Writing*.

The revised prototype assessment was administered to a small local group of students to determine the impact of the changes. Comments were candid and helpful. Reactions to the bell notification at strategic points in the assessment and to the two-speaker format at the top level of listening were favorable.

Copies of the prototype script (along with a reviewer's guide and test description) were given to two subject matter experts for review. Dr. Donovan Ochs was the listening expert; Dr. Anne DiPardo was the writing expert. Comments were submitted in written form and further discussed in personal meetings between the experts and the Work Keys development team. After analyzing the skills required at four levels of difficulty, Dr. Ochs concluded, "As a university professor engaged in the task of teaching communication skills I can assure you such a test will have immediate diagnostic uses as well as the potential for screening job applicants." He critiqued the entire test from directions through prompts and Scoring Guide. One suggestion he made was that examinees might be able to write more completely and accurately if they had more background information about the setting and the roles of the various people making the calls, receiving the messages, or otherwise mentioned in the messages. He also thought that interactions, such as the ability to ask questions, would be desirable.

Dr. DiPardo's comments about the *Writing* assessment indicated that the lower levels are clearly easier since the information load is less and rearrangement is seldom required. However, there is therefore also the tendency to write such messages in phrases rather than in complete sentences. At higher levels, the issues of deleting unnecessary information, ordering information, and conveying tone are potentially problematic. DiPardo also objected to the term "standard English" especially when used in conjunction with the word "correct" and strongly advised ACT to use terms such as "effective," "appropriate," or "clear" instead.

ACT's Director of the Office of Minority Student Development, Dr. Velma Tribble, also reviewed the assessment for fairness. She found the assessments to be fair and appropriate but advised the use of more names indicating the ethnic diversity of characters in the scripts.

B. Audiotape preparation

The next step in the development process involved the recruitment of script writers. Eight people (six female, two male) were selected and trained in a two hour session at ACT. Each was contracted to create 8 messages, 2 at each of four levels. These messages were edited by Work Keys development team members and then compiled into seven assessment forms, each with 12 messages. The audio production company was instructed that reading speed should be 130 words per minute; the messages to be read, the timing between message repetition, and the time allowed for writing messages were specified exactly; and the accents, genders, and ages of the readers to be used for particular messages were designated so that assessment forms would be consistent and comparable. Fifty copies of each audiotaped form were made.

C. Pretesting

School and business sites in Iowa, Michigan, Ohio, Tennessee, and Wisconsin administered the pretest materials following an administration pattern designed to provide randomly equivalent groups of examinees for each of the forms. A total of 3,319 examinees was assessed during March, 1992, each examinee completing one of seven pretest forms of the assessment. The number of examinees per form ranged from 341 to 703. Almost 95% of the examinees were students in grades 10 through 12, while the remainder were employees and community college students. Because the *Listening and Writing* assessment is presented via audiotape, spiraling of forms (using multiple forms of the same assessment in the same classroom) was not possible. However, participating sites were asked to randomly assign examinees to the various forms used at a specific site. In addition, two anchor prompts (identical messages used in all seven forms, in the same place in each form) were used to facilitate equating.

D. Scorer training

Pretest scoring was done by Measurement, Inc. (MI) of Durham, NC, under the general direction of the Work Keys development team. Using prototype responses, Work Keys staff prepared exemplars of messages at various score points for both the Listening and Writing

rubrics and spent two days at MI explaining the rubric and exemplars. MI staff then prepared additional exemplars, and hired, trained, qualified, and supervised their own scorers. MI divided Listening scorers into two groups, each responsible for a different set of messages, to reduce the information load. Writing scorers handled all the messages in a test booklet. Scorers were trained specifically in either the Listening or Writing rubric; no one scored both Listening and Writing. Training materials included a guide comprised of 22 sample messages illustrating specific score points, and six sets of ten sample messages, each used as practice/qualifying materials.

A similar process is now used by ACT; scoring was moved to the ACT scoring center in 1993. First, "range-finding" was conducted as a joint effort by staff from both Work Keys and the ACT Scoring Center. Exemplar responses were chosen to illustrate each score point, to serve as discussion examples for particular issues that might make assigning a score point more difficult, and to be used as training practice and qualifying sets. Justifications were written for each illustration (called an "anchor") and for each discussion exemplar explaining the reason for the assigned score. Updates, consisting of adding illustrative responses and strengthening the language of justifications, continue to be made at regular intervals.

Potential Work Keys scorers are initially screened through a resume review. Selected candidates are invited to complete a one-hour work sample. The work sample consists of assigning scores to sample messages based on the actual scoring rubric and exemplars. One may apply to do either Listening or Writing scoring, but not both, so that overlap in the evaluation of one skill with the other is eliminated.

"Table leaders" are trained first, after being selected for their previous experience in constructed response scoring and/or exceptional performance on the work sample, at a ratio of approximately one table leader to every eight to ten scorers. Table leaders assist in training the remaining scorers and score the more challenging responses. They also serve as resolution scorers when the two initial scorers assign ratings that differ by more than one score point.

Scorer training begins with an overview of Work Keys and the *Listening and Writing* assessment. Scorers are also provided with a general explanation of the scoring criteria. Next, Listening scorers are given a script of each message they will be scoring along with listings of primary and supportive details for each. Writing scorers are not given either the script or listings of content details since that information is irrelevant for their scoring. All scorers are shown skill-specific exemplars at each score point with detailed explanations given for each. Questions are taken at any time. Following instruction using the anchors, trainees sample score, analyze, and discuss sample responses. Then scorers complete a practice set of approximately 30 unscored responses, discussing any confusing issues as they practice scoring. Finally, qualifying sets are administered. Listening scorers complete five sets of 18 responses each; Writing scorers complete three sets. Newly-trained scorers are required to achieve 65% perfect agreement with the key and 90% perfect or adjacent (one score point higher or lower than the key) agreement in order to be hired. Monitoring of the percentage

of papers requiring resolution for each rater, "drift" toward either end of the scale, and monitoring of reading rate is done both electronically and by table leaders throughout the entire scoring process with retraining or release of scorers done as necessary.

E. Test administration

Administration procedures for all the Work Keys assessments are spelled out in the *Administrator's Manual*. They include specifics regarding security of materials, physical arrangements for testing, which materials ACT provides and which are to be provided at the local site, and what to do in the event of problems such as examinee illness or equipment failure. The setting is of particular importance in the administration of the *Listening and Writing* assessment since it is essential that examinees be able to hear the audiotape clearly and have adequate facilities for writing their responses. It is therefore recommended that no more than 25 to 35 examinees be tested in one room. Examinees have, on occasion, written in their answer booklets that they were unable to hear the tape, generally either due to ambient sound (testing of too many people at one time), use of inadequate equipment, or outside noise interference.

Listening and Writing has a constructed response format with examinees writing their responses on specially designated pages in the answer booklets. To make reading easier for scorers, examinees use pens with blue or black ink, not pencils. Unscored notes are made on the shaded left-hand pages; scored final responses are to be written on the lined, right-hand pages. ACT recommends that if other Work Keys assessments are also being given to examinees and the combined answer booklet is being used, the *Listening and Writing* assessment be given last. This is to prevent examinees from altering their responses during subsequent test sessions.

Accommodations including the use of a signer, scribe, special chair, table, special lighting, enhanced audio, or other physical accommodation, or a foreign language dictionary, or self-determined review of the audiotape under untimed conditions, can be made for examinees with special needs. A notation of "non-standard" conditions accompanies examinee scores if accommodations are provided.

Most directions to the examinees are both spoken on the audiotape and printed in the answer booklet so examinees can follow along with the audio. Directions emphasize that nothing on the notes pages is scored and everything on the message pages is scored. Thus, examinees may write in the language of their choice on the notes pages but must write their final messages in English. Likewise, drawings or vulgarities on the notes pages will be ignored, whereas they will detract from the examinee's score if placed on the message pages. The directions also emphasize the importance of keeping the messages business-appropriate, describe the aspects of the message which will affect scoring, and explain the proper use of the third person (many wrote as if the message was from them rather then from another).

In response to client feedback and statistics showing that some examinees stop writing, use inappropriate language, or otherwise express frustration and/or fatigue toward the end of the assessment, the directions have been altered in an attempt to encourage perseverance. For example, it is explicitly stated that the test consists of six messages and that each succeeding message is more complex.

F. Rating examinee responses

In an effort to keep the Listening and Writing skills scores as distinct as possible, the scoring for the Work Keys *Listening and Writing* assessment has been consistently and deliberately divided in two parts. Listening accuracy was initially scored analytically based on the number of pieces of information examinees recorded accurately, while Writing was scored holistically, focusing on the quality of the text produced.

After the prototype administrations, the final written messages were scored for Writing quality and the notes were scored for Listening accuracy. Directions read orally as well as printed in the booklet clearly explained this scoring procedure and gave further instructions to make the notes as complete as possible. However, scorer concerns about scoring only the notes when scoring Listening were soon reported. These included: (1) The written message sometimes contained information not present in the notes so the examinee obviously heard the information and simply did not note it; and (2) The notes were typically in a personal style, difficult to follow, and often did not convey any relationship between the pieces of information, so the examinee's understanding of the message could not be assessed.

To evaluate the effect of scoring the notes for Listening, a study was conducted using a sample of 78 examinees from the 1992 pretest administration. Each of the examinees' responses was judged separately based on the notes and on the written messages using the Listening scoring criteria. Analyses indicated that "In the [test] forms where differences between scores were exhibited, note-score was always lower than the message-score. In other words, using notes as a scoring basis produced a lower score than using written messages as a scoring basis." (Vansickle, 1992, p. 2). Vansickle indicated that notes were found to be less complete than written messages, that examinees could have been retaining the information auditorily rather than writing it down (even though they were asked to do the latter), and that raters could have been distracted by the quality of the response to be scored. "Because the written messages were more organized and polished than the notes, scores based on the former could have been inflated." (Ibid. p. 3). Based on the study results and rater comments, it was decided after pretesting to score the final written message for both Listening and Writing.

A more holistic approach to evaluating Listening also seemed desirable since preliminary scoring raised the following questions: what constitutes a piece of information; should all pieces of information have equal value (and, if not, who should judge which pieces have more value than others); and what percent of the information should the examinees have to capture in order to establish their mastery of a level?

The above issues were raised with the panels of experts (see Appendix section on skill scales) and, after considerable discussion, they advised Work Keys to return to the original process of scoring the final written message twice, once for Listening and once for Writing. The panels also helped develop the scoring criteria through a process of taking assessment messages themselves, looking at samples of examinee responses, ranking the responses, and then discussing reasons for their ratings. The resulting scoring criteria is intended to be a modified holistic model, with some specific details listed for each listening message to be used as scoring guides in addition to an overall evaluation of message completeness and accuracy. For Writing, some grammatical specifics were designated for each score point but the overall focus remained holistic. Minor wording modifications have been made over time. The current scoring rubrics are provided in the Appendix.

Two Work Keys staff meetings in March, 1992, one on test content and the other on process, focused on the analysis of the development of the assessment to date, and possible improvements in the development process. In the *Listening and Writing* assessment content meeting, issues such as clarity of directions, realism of messages, timing, and scoring were all addressed. In the process meeting, planned steps in the development process, actual occurrences, and changes to be made in the future were all discussed, and changes were recommended for the final form of the *Listening and Writing* assessment.

G. Scoring and scaling

Generalizability analyses (Brennan et al., 1995) prompted the decision to use two scorers and six prompts in the operational assessment. This approach was designed to shorten the total length of the assessment while maintaining a respectable level of reliability for each skill area. Assessment directions were revised to emphasize the need to include all of the information presented as well as to write in standard business English. Materials for the Work Keys job analysis (profiling) system were also developed to permit employers to use the Work Keys skill scale in identifying the Listening and Writing skill requirements of their jobs.

H. Final form

In the final form of the Work Keys *Listening and Writing* assessment, examinees are presented with a series of six audiotaped prompts at four graduated levels of complexity. The first prompt is at complexity Level 1, the second and third are at complexity Level 2, the fourth and fifth are at complexity Level 3, and the last is at complexity Level 4. Examinees are instructed to take notes as the prompts are presented, in the brief pauses between prompts, and during one repetition of each prompt. Following the repetition, the examinees are given a period of time to write out a message containing the information provided in the prompt. The audiotape controls the timing, thus it is uniform across administrations. The time allotted increases as the prompts become more complex, with the most complex prompt allowing six minutes of writing time. Examinees are permitted to review and work on previous prompts if they have time later in the test. Separate skill scales, scoring rubrics, and training materials

are used for Listening than are used for Writing, and different scorers are trained for each skill area.

Operational implementation of the *Listening and Writing* assessment occurred in 1993 and has been ongoing since. A number of evaluations of this unique assessment have taken place using both statistical and judgmental approaches.

III. Results of the operational *Listening and Writing* assessment to date

A. 1993 Wyoming study

A research study was conducted in Wyoming in 1993 using 12 test sites, 8 comprised of vocational education students receiving funding under the Carl Perkins Act, and 4 general education sites. Approximately 25 students were scheduled to participate at each site. The sites were selected to provide a diverse set of individuals with respect to gender, race and ethnicity, in both rural and urban settings. Three forms of the *Listening and Writing* assessment were included in the study, with the operational form, 10CC, used to report results to the examinees. Each examinee took two test forms, 10CC and either 11CC or 12CC in counterbalanced order. There are no common prompts across these forms.

To support generalizability analysis, 3 raters were used for each skill area for each booklet. Different raters were used for Listening than for Writing, and each test form (10CC, 11CC, and 12CC) had a different set of 3 raters as well, for a total of 18 raters used. Standard Work Keys Listening and Writing scoring rubrics were applied.

A total of 256 examinees actually completed the assigned pair of assessments. Forms 10CC and 11CC were taken by 167 examinees (Sample 1), and Forms 10CC and 12CC were taken by 89 examinees (Sample 2). No selection biases were noted. Although it had been attempted, counterbalancing appeared not to have been strictly followed in the test administration.

In the scoring, 1.7% of the Listening and 1.3% of the Writing ratings required resolution. Internal consistency reliability was .89 for Listening and .93 for Writing based on sets of 12 ratings (2 raters for 6 prompts). Correlations between Listening and Writing ratings averaged .55 on the same form and .54 across forms (the disattenuated correlations were .81 and .79 respectively). This suggests that rather than being caused by the common prompt, the relationship between Listening and Writing is inherent in the individual since prompts were not common across forms and the correlation remained about the same.

Three different designs and 2 different scoring approaches were used to conduct generalizability analyses of the assessment. The designs were: (1) person by (rater by task) nested in form, (2) person by rater by task, and (3) person by form. As used here, rating refers to the number assigned by a rater (one number for each skill for each prompt). Scoring refers to the summarization of that information for a single skill area over raters and prompts

in a number that can be reported to the examinee and others as the examinee's level score. The scoring approaches, used only in the third design (person by form), were the Guttman-based pattern scores used in operational scoring, and average scores (rounded and unrounded). Additional designs and results are provided by Gao (1996).

In the study based on the first design, person by (rater by task) nested in form, Writing had only two large variance components, the primary one being for persons accounting for 46% and 37% of the variance for Sample 1 and Sample 2, respectively. Person by task nested in form accounted for 29% and 32% of the variance. Task effects were small, indicating that the longer, more complex prompts did not constitute a more difficult writing task, and rater effects were negligible.

Listening had a different pattern with the largest variance component being for person by task nested in form, with 38% and 39% of the variance for the two samples. Task had the second largest components with 22% and 21%, with person effects accounting for 21% and 18% of the variance. Rater effects and their interaction with other components were very small for both skill areas.

Forms were found to be different in difficulty for Listening but not for Writing, indicating that equating activities would be needed for Listening. The presence of person by form interactions for Listening, however, could not be addressed by this means.

In the second design, person by rater by task, Listening continued to show large task variance, and both skill areas showed person by task interactions (prompt specific variance) although this continued to be larger for Listening than for Writing. Again, rater effects were small.

For the third design, person by form, the same types of effects reported in the first two designs were noted. In addition, two different scoring methods were investigated. Variance component estimates for person by form interactions for the mean ratings (both rounded and unrounded) were smaller than for pattern scores, leading to a reduction in measurement error. Decision studies (based on absolute rather than relative error variances) were conducted for each of the designs. The results indicated that adding raters, beyond the two in operational use, would not improve measurement precision very much, but that adding more tasks and/or forms would. This effect was greater for Listening than for Writing.

Dependability coefficients estimated from the first design, assuming one form, two raters, and six tasks (the operational configuration), were .81 for Writing and .56 for Listening, for Sample 1, and .72 for Writing and .51 for Listening, for Sample 2. In the second design, assuming two raters and six tasks, the dependability coefficient ranged from .76 to .86 for Writing and from .60 to .66 for Listening. Finally, in the person by forms design, it was found that use of mean scores would provide noticeably higher levels of generalizability than did the current pattern scoring.

BEST COPY AVAILABLE

Gao concluded that "the use of six prompts and two raters in the Work Keys *Listening and Writing* assessment leads to better generalizability than other performance assessments with fewer tasks and/or raters" (p. 25). She cautions, however, that the different task ordering of examinees on the different forms, due to large task-sampling variability (especially for Listening), suggests that conventional equating methods may not be entirely satisfactory. The findings with respect to scoring methods also suggest that alternative approaches for converting raw scores to level scores be explored.

Multivariate generalizability studies were also conducted using these data. A distinguishing feature of multivariate generalizability theory is that it involves not only variance components but also covariance components. A multivariate generalizability analysis facilitates the examination of relationships between universe scores on different measures (variables) such as the *Listening and Writing* assessment as well as possible correlated errors between these measures. For the generalizability (G) studies, persons and tasks are considered "jointly sampled" in the sense that the G studies employ the same persons and the same tasks (prompts) for both Listening and Writing; but raters are assumed to be independently sampled because different raters scored the responses for Listening and for Writing. Consequently, covariances that involve persons or tasks, but not raters, are not necessarily zero. However, matrices that involve the rater effect, r, do have zero covariance. Variance-covariance matrices for the effects of person, task, and person by task interaction, are presented in Table 1. Each correlation in Table 1 indicates how strongly the respective effects are correlated. A high disattenuated correlation for Listening and Writing scores indicates that the two skills are related. The results in Table 1 suggest that the universe scores between Listening and Writing are substantially correlated for both Form 10CC (.81) and Form 11CC (.93). Interestingly, even if the tasks on the *Listening and Writing* assessment were not jointly sampled (e.g., Form 10CC *Listening* with Form 11CC *Writing*), the disattenuated correlation coefficients are about .75.

Table 1

Multivariate G study Variance-Covariance Components for Listening (L) and Writing (W)

Source	Variance		Covariance	Correlation
	Listening	Writing		
		Form 10CC		
Person	0.27	0.43	0.28	0.81
Rater	0.01	0.00	0.00	0.00
Task	0.23	0.02	0.04	0.58
pr	0.01	0.02	0.00	0.00
pt	0.50	0.24	0.04	0.13
rt	0.00	0.01	0.00	0.00
prt	0.15	0.18	0.00	0.00
		Form 11CC		
Person	0.43	0.35	0.36	0.93
Rater	0.00	0.01	0.00	0.00
Task	0.02	0.01	0.01	0.10
pr	0.02	0.02	0.00	0.00
pt	0.24	0.23	0.05	0.20
tr	0.01	0.00	0.00	0.00
prt	0.18	0.12	0.00	0.00

The covariance component is the covariance between task mean scores for Listening and Writing, where the mean is the expected value taken over all persons in the population and all raters in the universe of admissible observations. As indicated in Table 1, the covariance is relatively small for Form 10CC but relatively large for Form 11CC compared to the variance components of the task effect. The correlation coefficient of the task effect is moderate for Form 10CC but high for Form 11CC. These results suggest that there is a moderate to high relationship between Listening and Writing mean scores based on the same task. The covariance component is small relative to the variance components. This suggests that examinees tend to be rank ordered differently by their performance on different tasks but that the differences in the rank orderings are not very consistent across Listening and Writing. The

variance and covariance components in Table 1 are the principal results of the multivariate G study.

B. 1994 Operational data studies

Based on the performance of 7,097 examinees assessed in the Spring of 1994, scoring consistency remained high for both skills. For Listening skills, 74% percent of the 42,582 responses (7,097 examinees x 6 prompts) were assigned the same score by both raters. Raters gave another 22.8% of the responses adjacent scores on the five-point scale. The remaining 3.2% of the responses received nonadjacent scores which were resolved by a third scorer. For Writing skills, raters agreed exactly on 62.6% of the responses rated and assigned an additional 35.1% of the responses adjacent ratings. The remaining 2.3% of the responses required resolution. Correlations between the ratings assigned by the first and second raters for Listening and Writing based on the data from 7,219 examinees are presented in Table 2.

Table 2

Correlations Between First and Second Raters' Scores

Prompt	Listening	Writing
1	.81	.77
2	.81	.76
3	.79	.75
4	.78	.78
5	.80	.80
6	.82	.81

An "end of booklet" effect was apparent based on the percentage of non-scorable responses. For the first three prompts, non scorable responses averaged between 1.6% and 1.8% of the examinee group for both Listening and Writing. However for the fourth prompt this percentage averaged about 2.6 for Listening and 2.3 for Writing; for the fifth prompt it averaged 3.4% for Listening and 3.2% for Writing; and for the last prompt, it averaged 6.8% for Listening and 6.1% for Writing. Although a six to seven percent non-completion rate for a test is not necessarily a problem, in this case the test was paced by the audio-tape, ensuring that every examinee had an opportunity to address each prompt. The fact that this proportion of individuals did not address the last prompts (which were the most difficult) or did so in a non-scorable manner (generally with profanity or other off-topic response), suggested fatigue, an unwillingness to attempt a more difficult task, or both (Vansickle, 1995).

Both factor analysis and generalizability procedures revealed the end of booklet effect as well as significant prompt-specific effects for Listening but not for Writing. Factor analysis showed a primary factor accounting for 43.9% of the variance for Listening with the second factor accounting for 12.4% of the variance. Five factors had eigenvalues larger than 1 and together accounted for 83% of the variance. The first factor was defined by the last two prompts (5 and 6). The remaining factors were prompt specific with prompt 1 defining the second factor followed by prompts 2, 3, and 4, in that order. Writing had a single primary factor with an eigenvalue of 7.45 accounting for 62% of the variance. No other writing factor had an eigenvalue greater than 1.

A generalizability analysis showed similar results. Variance components for Form 10CC assuming six tasks and two raters per task are shown in Table 3.

Table 3

Variance Components for Form 10CC with 6 Tasks and 2 Raters

	Listening	Writing
Persons (p)	.207	.441
Tasks (t)	.202	.070
Raters (r)	.000	.000
pt	.499	.289
pr	.006	.029
tr	.000	.000
ptr	.108	.146

Note that the scores on which these analyses are based were computed as the average ratings given over the six prompts by two raters per prompt rather than as the categorical level scores which were reported to examinees. This approach was chosen to reduce the impact on the analysis of the small number of score points on the level scale, and to explore the possibility that the use of a reporting scale based on score averages instead of on the original pattern scoring, would improve the reliability of the reported scores.

The correlation between Listening and Writing scores for Form 10CC using this data set was .56 based on the average ratings treated as continuous variables (i.e. raw scores). Table 4 provides cross tabulations of categorical levels, based on rounding the average scores. Ratings ending in .5 were rounded upward.

Table 4

Percentage of Examinees by Listening Level and Writing Level (N = 7,097)

Listening Average Rating	Writing Average Rating						
	0	1	2	3	4	5	Listening Percent
0	0.15	0.15					0.31
1	0.10	1.04	0.70	0.07			1.92
2		0.62	9.30	11.74	2.96	0.01	24.63
3		0.01	6.06	31.94	27.12	0.39	65.53
4			0.08	1.99	5.19	0.28	7.54
5				0.01	0.06		0.07
Writing Percent	0.25	1.83	16.15	45.75	35.32	0.69	

C. Conceptual analysis

From a conceptual standpoint, the relationship between the Listening and Writing skills as measured on the Work Keys *Listening and Writing* assessment is not symmetrical. An examinee must listen well enough to provide an on-topic response for the Writing score. This is accomplished at a Listening level of 1. However, an examinee must also write well enough for the rater to be able to accurately determine what was heard. This is accomplished at a Writing level of 2. Based on this consideration, ACT recommends that whenever the *Listening and Writing* assessment is administered, both Listening and Writing be scored, and that Writing scores of individuals who score below 1 on Listening, and Listening scores of individuals who score below 2 on Writing, be considered invalid.

Additional considerations may apply at the top of the scale. Cross-tabulations and rater comments indicate the possibility of negative interference at the top of the score scale. Individuals who record all of the information exactly as it was given (scoring at Level 5 on Listening) may not have reworded the information sufficiently to generate a Level 5 written message. Conversely, those who have written very fluently and at a very high level of

competency (scoring at Level 5 on Writing) may have reworded enough or dropped out enough details that they do not relate all of the information accurately, thus scoring below Level 5 on Listening. Alternatively, it may simply be very difficult to conjunctively meet the very highest standards on both skill scales.

In August of 1994, ACT conducted a joint research study with an insurance company to investigate the appropriateness of the Writing score from the Work Keys *Listening and Writing* assessment for measuring the writing skills of employees responding to written stimuli. The Work Keys *Listening and Writing* assessment was administered to 186 customer service representatives. Based on the Writing scores, the sample was stratified and 74 of the representatives were selected to reflect the percentage of customer service representatives who had scored at each of the writing levels.

The company provided ACT with copies of two letters written by each of these 74 individuals, and each letter was scored using the same procedures and scoring criteria as those used for scoring written responses to the Work Keys *Writing* assessment. The correlation between the individuals' scores on the *Writing* assessment and on the two letters was .89. This approaches the limit of correlation possible given the reliability of the two letter samples, which was about .73. Indeed, the relationship between the two letters written by the same individual was lower than that between the letters and the Writing test score. The results of this study indicated that it was appropriate to use the Work Keys *Writing* assessment, with its audio stimulus, to assess these customer service representatives' writing skills, even though the customer service representatives generally respond to written stimuli in their workplace.

Concerns were raised in job profiling about the appropriateness of the Listening score from the *Listening and Writing* assessment for jobs in which responses to listening do not usually involve writing. In many factory and other settings, an employee responds to a verbal direction by behaving in a particular way, by "doing something." The issue was whether it was appropriate to require writing in response to the auditory stimulus as part of the assessment when writing is not required on the job.

It was clear from a legal standpoint that it would not be appropriate to require writing skills for assessment purposes if the job did not require writing skills. Even where the individual possessed sufficient writing skills to be appropriately assessed, allowing writing to become a significant part of the assessment criterion would not meet the EEOC requirements (reference) for job relevance.

Beyond that, it appeared that jobs in which verbal information is not written down use messages that are different from jobs in which writing down the information is expected. When a message is given to someone who is writing, it can be short with only a little information or it can be quite lengthy with many pieces of information. The range is similar to that included in the current *Listening and Writing* assessment. However, when the message will not be written, it is typically short with only a few pieces of information. If lengthier

BEST COPY AVAILABLE

information must be given, it will generally be given in written form, or if auditory, will be broken up into a series of shorter components. Thus, the messages used in a non-writing environment are similar in complexity to those at Levels 1 and 2 on the *Listening and Writing* assessment (prompts 1 to 3), but are unlike those at complexity Levels 3 and 4 (prompts 4 to 6). Accurate listening remains critical for these jobs (the job profiles at a Level 4 or 5 of the Listening scale) but the messages listened to are short and of limited complexity.

To investigate the possibility of developing a listening assessment for this type of situation, ACT modeled a test comprised of prompts 1 through 3 on Forms 10CC and 11CC using an IRT approach (Wang, 1995b). Sample sizes for these forms were 7,097 and 2,035 respectively. A partial credit polytomous model was fitted to each of the data sets using FACETS (Linacre, 1989). To meet the requirements of the software, a 50% sample of each examinee pool was used for calibration. The averaged standard deviation was used as the standard deviation for theta. The expected marginal level score distribution, classification indices and reliability indices were computed using methods described by Wang (1995b).

Expected score distributions for the reconfigured form showed higher frequencies of examinees at the upper end of the score distribution than for either 10CC or 11CC. This effect was more pronounced for Listening than for Writing. The reliability of the reconfigured form (.73 for Listening and .80 for Writing) was between those for 10CC and 11CC, indicating that the reconfigured instrument might work well for jobs with limited or no writing requirements (Wang, 1995a).

IV. Conclusion

From the initial specifications development, through pilot testing and pretesting, and into operational use, ACT has carefully researched and monitored the Work Keys *Listening and Writing* assessment. As a result, much has been learned about these skills, both as they occur independently and as the assessment of one affects the assessment of the other in this joint format. It has become clear that Listening and Writing skills are fairly highly correlated, independent of the assessment format. It is also apparent that except for a tendency of examinees not to complete the assessment attentively due to its length, the complexity of the last prompts, or both, the use of the audio prompt has little impact on the assessment of writing skills. The ability to demonstrate skill in constructing technically correct English (grammar, syntax, spelling, punctuation and the like) with minimal interference from examinee background and cultural influences, seems well supported in this assessment format.

Although prompt-specific variance has lowered the internal consistency reliability of the Listening score of the assessment, despite very high inter-rater reliability, this may not be the result of the constructed response format. Rather, it seems likely that there are other variables, such as the examinees' previous experience with the specific context and vocabulary, or there may be features similar to modes of discourse in writing, that have had an effect on the Listening scores. Additional research will be required on this point.

A major advantage gained by separately scoring the Listening and Writing skills measured by this assessment has been the resulting possibility of examining the impact of one on the other. This impact would, of course, exist with or without the separate scores. Since all assessments require the simultaneous demonstration of multiple skills to some degree, and since assessments using a written response format may place significant demands on examinee writing skills, it is important to evaluate the individual and joint effects of the multiple skills on the performance of individuals on the assessment even if they are not scored separately for reporting purposes. This study has demonstrated some of the approaches that can be used in this type of research, and has demonstrated the wide variety of potential effects that must be considered.

References

American College Testing. (1995). Work Keys Administrator's Manual. Iowa City, IA: Author.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys Listening and Writing Tests. Educational and Psychological Measurement 55,(2) 157-176.

Gao, X. (in press). Sampling variability and generalizability of Work Keys Listening and Writing scores (ACT Research Report). Iowa City, IA: American College Testing.

Linacre, J. M. (1989). Many-facet Rasch Measurement. Chicago: MESA Press.

U.S. Equal Employment Opportunity Commission (1978). Uniform guidelines on employment selection procedures. Washington, DC: Author.

Vansickle, T. R. (1992, August). Work Keys: Developing a usable scale for multi-level, criterion-referenced assessments. In J. D. West (Chair) Work Keys supporting the transition from school to work. A symposium presented to the annual meeting of the American Psychological Association, Washington, DC.

Wang, T. (1995a). IRT-based analysis of the reconfigured Work Keys Listening and Writing tests. Unpublished manuscript, American College Testing, Iowa City, IA.

Wang, T. (1995b). An IRT-based analysis of the Work Keys Listening and Writing tests. Unpublished manuscript, American College Testing, Iowa City, IA.

Appendix

Skill Scales

The skill scale descriptions and scoring rubric were developed and refined by the Work Keys development team in conjunction with two panels of experts, one panel focusing on listening skills and the other on writing skills. Panel members consisted of one business and one education representative from each of five states, with the listening panel convening April 1-2, 1992, and the writing panel meeting April 7-8, 1992, both in Iowa City. Additional refinement of the skill scale descriptions occurred as the listening and writing skills were profiled in various job situations. The current descriptions are as follows:

Listening Skill Scale

Level 1: People with Level 1 skills can write down a small amount of useful information based on a spoken communication. This written information may give clues to the gist of the situation or provide sources of further information, but it does not include enough information to give the receiver a correct understanding of the situation described in the message.

Level 2: People with Level 2 skills can correctly write down the basic ideas of a spoken message, giving a fair amount of useful information, but may miss details or incorrectly record some of the information.

Level 3: People with Level 3 skills can listen to a spoken communication and record messages that are basically correct, but they may miss one or two pieces of important information.

Level 4: People with Level 4 skills can correctly record all the important information and the relationships among pieces of information, in addition to accurately conveying the central idea of a spoken message. However, they may miss or incorrectly record some details, or include irrelevant information.

Level 5: People with Level 5 skills can correctly record all the important information and the relationships among pieces of information from a spoken message. They also use supporting details to convey insight into the particular situation the message involves.

Writing Skill Scale

Level 1: People with Level 1 skills can write messages in English. A large number of major grammatical, punctuation, spelling, and/or other mechanical errors make the messages very unclear and inconsistent with standard business English.

Level 2: People with Level 2 skills can write messages that are generally understandable. Many errors in grammar, punctuation, and/or sentence structure make understanding these messages somewhat difficult.

Level 3: People with Level 3 skills can write clear messages which include some incomplete sentences and/or errors in grammar and punctuation.

Level 4: People with Level 4 skills can write messages that are clear and generally consistent with standard business English. Such messages contain complete sentences but may include a few minor errors in grammar and punctuation, and/or the writing style may lack clear organization and appropriate transitions.

Level 5: People with Level 5 skills can write messages that are clear and highly consistent with standard business English. They can use good sentence structure without mechanical errors, and a smooth, logical style.

Note: Examinees with limited writing skills may be unable to express themselves well enough in writing to accurately demonstrate their listening skills. Likewise, examinees with extremely limited listening skills may be unable to produce a response that is sufficiently on topic to accurately demonstrate their writing skills.

Scoring Rubrics

In the original scoring rubrics "primary" details were labeled "critical/essential" and "supportive" details were described as "non-critical/non-essential" but these designations proved problematic particularly in job profiling since subject matter experts routinely maintained that virtually everything they heard on the job was critical and almost nothing was non-critical. This resulted in very high expectations for listening scores. Job profiling trainees have also raised some job-specific distinctions concerning which information is labeled "primary" or "supportive", e.g., where the customer purchased the item to be repaired may be of greater significance to one business than to another. Other job-specific issues include how to score implied information (i.e., whether to give credit for such or consider it missing), definition of "business acceptable" (e.g., some businesses are far more casual about the use of slang than are others), and detail level required (e.g., how significant is the difference between "on Friday" and "by Friday").

Listening Scoring Rubric

Score 5: All primary and supportive details are present and correct, including all relationships among details.

Score 4: Response is correct in that all primary details and relationships among details are given and correct; may be missing supportive details or have incorrect supportive details that do not interfere with accurate communication.

Score 3: Response is substantially correct; all the primary details present are correct and relationships among them are correct; may be missing a few primary details.

Score 2: Some pertinent details; may have incorrect primary details, but sketch of the situation is correct.

Score 1: Minimal pertinent information; enough details to provide clues as to gist of situation OR source of further information.

Score 0: No meaningful information, or totally inaccurate information.

Writing Scoring Rubric

Score 5: Conveys message clearly; highly appropriate for the business setting of the prompt; no mechanical errors; good sentence structure; smooth and logical style.

Score 4: Conveys message clearly; may have a few minor mechanical errors that do not interfere with comprehension; good sentence structure (e.g., all sentences are complete); adequate style (sentences may be somewhat choppy; overall message may not be completely smooth or logical).

Score 3: Conveys message clearly; some mechanical errors (e.g. problems with spelling, punctuation, etc. that do not interfere with comprehension); adequate sentence structure (e.g., most sentences are complete).

Score 2: Conveys message adequately; many mechanical errors (problems with spelling, punctuation, etc. interfere with comprehension); weak sentence structure (incomplete sentences or poorly structured sentences with comma splices, fused, etc.)

Score 1: Conveys message inadequately; gross mechanical errors (problems with spelling, punctuation, etc.) may make deciphering difficult, overall lack of proper sentence structure (may be very difficult to decipher).

Score 0: An attempt is made at the message, but the response is completely garbled, with no recognizable sentence structure; one or more comments within the response are grossly inappropriate (e.g., swear words, threats); message is off-topic; page is blank; or message is complete illegible.

Note: While writing style and mechanics do not affect the *Listening* score, examinees with limited writing skills may be unable to express themselves well enough in writing to receive a valid *Listening* score. Likewise, examinees with extremely limited listening skills may be unable to produce a response that is sufficiently on topic to receive a valid *Writing* score.